УДК 004.934

Н.Б. Васильєва, Д.Я. Федорин
Міжнародний науково-навчальний центр інформаційних технологій та систем,
м. Київ, Україна
03680, Україна, м. Київ, просп. Академіка Глушкова, 40

Проблеми створення систем розпізнавання мовлення для різних комп'ютерних платформ

N.B. Vasylieva, D.Ja. Fedoryn
International Research and Training Center of Information Technologies and Systems, Kiev, Ukraine
40 prospekt Akademika Hlushkova, Kyiv, 03680, Ukraine

Problems of Cross-platform Speech Recognition System Creation

Н.Б. Васильева, Д.Я. Федорин
Международный научно-обучающий центр информационных технологий и систем,
г. Киев, Украина
03680, Україна, г. Киев, пр. Академика Глушкова, 40

Проблемы создания систем распознавания речи для различных компьютерных платформ

Розглядаються проблеми, пов'язані з побудовою системи розпізнавання мовлення на різних обчислювальних платформах. Особлива увага приділяється формуванню бази даних і знань акустичного, фонетичного та лексичного рівнів. Моделюється зв'язок акустичної та лінгвістичної компонент системи розпізнавання мовленнєвого сигналу, досліджується ефективність вибору мовленнєвих елементів та застосовуються методи обмеження порядку їх слідування. Описуються особливості реалізації системи розпізнавання на архітектурі мікропроцесорів ЦОС, враховуючи можливість віддаленої обробки мовленнєвого сигналу.
Ключові слова: лінгвістична модель, акустична модель, крос-платформова система, розпізнавання мовлення

The problems associated with building a speech recognition system on different computing platforms are considered. Particular attention is given to the data and knowledge base forming for acoustic, phonetic and lexical levels. Relation between speech recognition acoustic and linguistic components is being modeled as well as spoken element selection has been investigated and element order constraining methods are applied. Aspects of decoder implementation on the DSP microprocessor architecture including the possibility of speech signal remote processing are described.
Key words: linguistic model, acoustic model, cross-platform systems, speech recognition.

Рассматриваются проблемы, связанные с созданием систем распознавания речи на различных вычислительных платформах. Особое внимание уделяется формированию базы данных и знаний акустического, фонетического и лексического уровней. Моделируется связь акустической и лингвистической компонент системы распознавания речевого сигнала, исследуется эффективность выбора речевых элементов и применяются методы ограничения порядка их следования. Описываются особенности реализации системы распознавания на архитектуре микропроцессоров ЦОС, включительно с возможностью удаленной обработки речевого сигнала.
Ключевые слова: лингвистическая модель, акустическая модель, кросс-платформенная система, распознавание речи.

Вступ

Розпізнавання злитого мовлення в реальному часі дозволяє вирішувати широкий спектр прикладних задач у різноманітних областях людського життя. Аналіз патентів комерційних фірм і публікацій відомих наукових центрів світу показує, що останнім часом з’явилося багато програмних засобів диктування на ПК, а також мережні сервіси, які дозволяють усно формувати пошукові запити або диктувати листи електронної пошти. Всі найбільш продуктивні системи реалізують генеративну модель аналізу, розпізнавання та розуміння мовленнєвого сигналу в тій або іншій модифікації [1-3].

Ефективність застосувань системи розпізнавання мовлення залежить від оцінювання багатьох параметрів, а це досі не достатньо вивчено. Залишаються відкритими питання вибору елементів розпізнавання на різних рівнях та взаємозв’язку рівнів системи розпізнавання.

Реалізація алгоритмів розпізнавання та синтезу мовлення в портативних пристроях є надзвичайно актуальною проблемою. Насамперед, це стосується алгоритму розпізнавання великих словників, тобто пофонемного розпізнавання ізольованих слів, причому кількість слів у словнику, які система може розпізнати, складає 1000 елементів та більше.

Найбільш актуальним є вирішення задач, пов’язаних з розпізнаванням злитого спонтанного мовлення та синтезом природною мовою довільного тексту. Це дало би змогу керувати голосом портативними пристроями, перекладати сказане іншими мовами, здійснювати голосовий пошук, розробляти діалогові системи тощо.

У залежності від місця, де відбувається перетворення «вимовлена фраза – текст» та «текст – вимовлена фраза», програма розпізнавання та синтезу мови поділяється на ізольовані (*client-side*), клієнт-серверні (*server-side*) та гібридні (*hybrid*).

В ізольованих системах перетворення відбувається безпосередньо на мобільному пристрої. У клієнт-серверних системах мобільний пристрій використовується тільки для введення інформації з подальшою її передачею по мережі на сервер для обробки та отримання від сервера відповіді розпізнавання або синтезованої фрази.

Гібридні системи поєднують в собі функціональність ізольованих і клієнт-серверних: при наявності доступу до мережі вони використовують для перетворення сервер, при недоступності мережі – працюють як ізольована система.

Прикладом реалізації ізольованої системи може бути система *CeedVocal* [4], прикладом клієнт-серверної системи є загальновідома *Siri* [5], прикладом гібридної – *VoCon Hybrid* [6]. Кожний із підходів має свої переваги та недоліки. Ізольована система обмежена швидкодією та розміром доступної оперативної пам’яті сучасних мобільних систем, що в свою чергу накладає обмеження на розмір словника і збільшує час відповіді застосування.

Клієнт-серверна технологія не має цих обмежень, але потребує для своєї роботи постійного підключення до глобальної мережі. Гібридна технологія, маючи, по суті, властивості двох попередніх технологій в одній системі, є найбільш гнучкою.

Далі ми розглянемо загальну структуру розпізнавання мовлення, проаналізуємо ефективність її компонент окремо та разом, а також розглянемо особливості адаптації мовленнєвих систем до різних, в особливості, до портативних платформ.

Загальна структура розпізнавання мовлення

Вхідний мовленнєвий сигнал перетворюється в послідовність акустичних векторів-ознак $\mathbf{Y}_{1:T} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T)$ в результаті постпроцесингу. Потім декодер намагається

відшукати послідовність мовленнєвих сегментів, заданих символами $\boldsymbol{w}_{1:L} = (w_1, w_2, ..., w_L)$, яка найбільш ймовірно відповідає $\boldsymbol{Y}$**,** яка спостерігається:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} P(\mathbf{w} \mid \mathbf{Y}) \cong \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{Y} \mid \mathbf{w}) P(\mathbf{w}). \tag{1}$$

Еквівалентність правої частини виразу, що випливає із застосування правила Байєса, представляє базове формулювання генеративної моделі розпізнавання мовлення. Акустична – $p(\boldsymbol{Y} \mid \boldsymbol{w})$ та лінгвістична – $P(\boldsymbol{w})$ – складові генеративної моделі, описуються кожна своїми стохастичними породжувальними граматиками.

Акустична модель кожного зі слів w формується в результаті композиції моделей базових мовленнєвих елементів, тобто фонем, які складають фонемну транскрипцію слова $\mathbf{q}_{1:K_w}^{(w)} = (q_1, q_2, ... q_{K_w})$. Для моделювання екстралінгвістичних явищ, властивих спонтанному мовленню, в алфавіт базових елементів, додатково до фонем і фонем-пауз, вводяться символи, які відображають неінформативні звуки.

Загальноприйняті системи пофонемного розпізнавання оперують алфавітом фонем, контекстно-залежних або контекстно-незалежних, з яких будуються мовленнєві образи слів. Вже на послідовності слів накладаються обмеження шляхом введення лінгвістичної моделі на основі породжувальних граматик або статистичної моделі, враховуючи контексти слів.

На рис. 1 зображено загальну структуру системи автоматичного розпізнавання злитого мовлення, яка є спільною для ізольованих, клієнт-серверних і гібридних технологій.

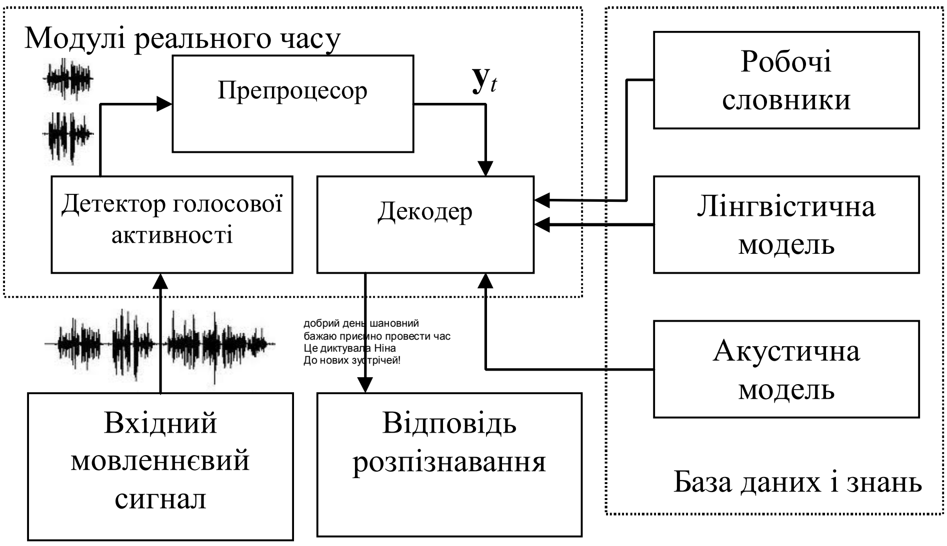


Рисунок 1 – Структура автоматичного розпізнавача злитого мовлення

До *модулів реального часу* надходить *мовленнєвий сигнал* через одне із доступних джерел: мікрофон, файлова система, звуковий запит з віддаленого пристрою тощо. При проходженні через *детектор голосової активності* сигнал розбивається на сегменти за ознакою спостереження голосового введення. Використовуються прості акустичні ознаки в амплітудно-часовому просторі на основі поточної амплітуди та кількості переходів через нуль. Модуль *препроцесора* переводить сигнал у простір первинних векторів-ознак. При цьому використовується мел-кепстральне перетворення з відніманням середнього значення. *Декодер* проводить порівняння вхідного мовленнєвого сегмента з гіпотезами еталонного сигналу допустимих послідовностей слів із *робочих словників*, застосовуючи деяку обережну стратегію відкидання малоперспективних гіпотез. Для цього використовуються дані з *акустичної* та *лінгвістичної моделей*. Послідовність слів,

за якою генерується еталонний сигнал, найбільш схожий на вхідний сигнал, оголошується відповіддю розпізнавання.

В ізольованих системах розпізнавання мовлення модулі реального часу та всі компоненти бази даних і знань знаходяться безпосередньо на портативному пристрої. Клієнт-серверні системи передбачають розміщення детектору голосової активності на пристрої, у той час як препроцесор може перебувати як на серверній, так і на клієнтській частині. Декодер із базами даних і знань перебуває на сервері. У гібридних системах розташування як модулів, так і компонент бази даних і знань варіюється.

Формування баз даних та знань для систем розпізнавання мовлення

База даних та знань для системи розпізнавання мовлення включає робочі словники, а також акустичну та лінгвістичну моделі.

У робочому словнику містяться варіанти вимовляння для кожного зі слів лексикону, який припускається. Графемно-фонемні перетворення потрібні для формування словників вимовляння при оцінці параметрів акустичної моделі та композиції акустичних моделей слів на етапі декодування. Використана система багатозначного транскрибування орфографічних текстів використовує кінцевий автомат, який передбачає можливість таблично задавати контекстно-залежні правила перетворення одних узагальнених послідовностей символів на інші [7]. Застосування багатьох правил дозволяє генерувати одразу декілька варіантів транскрипції одного і того ж слова, або генерувати потрібний варіант із декількох можливих, наприклад, описуючи спонтанне мовлення диктора або групи дикторів.

Можливість генерувати одразу декілька варіантів транскрипції одного і того ж слова дозволяє продемонструвати в словнику варіантність вимови найбільш частотних українських слів, редукцію та розтягнення слів під час швидкого темпу мовлення, нечітку вимову та подібні явища нарівні з літературним варіантом вимовляння. Також система транскрибування дозволяє генерувати транскрипції для таких специфічних підсловників, як суржик, соціальні та територіальні діалекти, абревіатури тощо.

Зв’язок словника з акустичною та лінгвістичною моделями здійснюється за ідентифікатором (іменем), доповненому ймовірністю приналежності до кластера слів для моделей, основаних на класах слів, які дають змогу суттєво зменшити обсяги лінгвістичної моделі [8].

Параметри акустичної моделі оцінюються на основі мовленнєвого корпусу, що складається зі структурованої множини мовленнєвих фрагментів, текстового опису цих фрагментів, а також інструментарію для оперування всією множиною даних корпусу.

Одним зі способів формування мовленнєвого корпусу є запис диктора, який зачитує деякий текст, в якому міститься все фонетичне розмаїття українського мовлення. Це дає змогу уникнути етапу ручного транскрибування та сегментування, а також одночасно формувати тестовий корпус (ТК), який відповідає досліджуваній предметній області. Формування такого тексту відбувається на основі електронних текстів, що знаходяться у вільному доступі в Інтернеті.

У процесі формування тексту навчальної вибірки (НВ) *злитого мовлення* проводилося перетворення чисел, символів і скорочень на послідовності графем, графеми перетворювалися на фонеми [7] з наступним використанням процедури, що дає змогу досягнути суттєвого скорочення тексту НВ без втрати фонемного розмаїття [9]. На етапі обробки тестового корпусу і формування НВ розглядалися фонеми-трифони як базові

мовленнєві образи, оскільки вони мають регулярну структуру і дають можливість моделювати фонемне різноманіття, враховуючи правий та лівий звукові контексти.

Для формування НВ *ізольованих слів* використовувалися частотний словник української мови та лексикографічна система УМІФ. Кількість фонем-трифонів, що належать обом словниковим вибіркам, складає приблизно 15 тис. елементів. При цьому 12 тис. фонем-трифонів належать тільки НВ на основі словника УМІФ, а 3 тис. – тільки НВ частотного словника. Обсяг словника НВ ізольованих слів склав 13 тис. слів (біля 12 годин запису).

Для перевірки ефективності використання мовленнєвих образів, тобто фонем, відкритих складів та складів, отриманих за правилами складоподілу, формувалися тексти контрольної вибірки (КВ) злитого мовлення і проводилася процедура запису в умовах, аналогічних запису НВ.

Перший спосіб вибору тексту КВ оснований на тому, щоб перевірити розпізнавання часто вживаних слів, речень, фраз, тобто сформувати КВ за частотою фонем-трифонів – *«частотну» КВ*. Отримана КВ містить три з половиною години запису та дозволяє відслідкувати помилки при розпізнаванні фонем у всіх типових контекстах. Обсяг словника складає 3 тис. слів. Загальна кількість реалізацій слів – приблизно 9 тис.

Другий спосіб полягає в формуванні КВ випадковим чином із тих самих текстів, з яких вибирався текст НВ, але із забороною вибору тих речень, які увійшли до НВ.

Отримана *«випадкова» КВ* містить чотири з половиною години запису та є найбільш типовою для вибору предметної області, тобто помилка розпізнавання в ній буде мати найбільш характерне значення вибраної предметної області.

Обсяг словника складає 10 тис. слів. Загальна кількість реалізацій слів – приблизно 23 тис.

Остання КВ вибиралася із текстів, які не використовувалися ні для формування попередніх КВ, ні для НВ. Для цього із сайту україномовної Вікіпедії випадковим чином вибрано 100 МБ текстів – *КВ «Вікіпедія»* (три години запису). Обсяг словника складає більше 7 тис. слів. Загальна кількість реалізацій слів – 16 тис.

Збільшення ефективності акустичної моделі

В якості композитного мовленнєвого елемента розпізнавання за допомогою декодерів *HTK* та *Julius* [2], [10] були взяті фонеми (всього 59), відкриті склади (всього 7 270) та склади, поділені за правилами українського складоподілу (всього 10 200).

Були проведені експерименти з дослідження впливу змісту НВ акустичної моделі на розпізнавання. Розглядалися такі варіанти акустичної моделі розпізнавання: модель, побудована тільки на злитому мовлені; модель, яка об’єднує злите мовлення та ізольовані слова; модель, яка не враховує або враховує лише частково наголошеність голосних.

В експериментальних дослідженнях оцінювалися показники фонемної помилки (*PER – Phoneme Error Rate*), що відображають відношення між різницею правильно розпізнаваних фонем і помилкових вставок до загальної кількості фонем.

У табл. 1 представлені результати фонемної помилки розпізнавання при використанні різних акустичних моделей.

Порівнюючи результати, наведені в табл. 1, бачимо, що використання акустичної бази НВ ізольованих слів на додаток до НВ злитого мовлення приводить до зменшення фонемної помилки.

Це можна пояснити тим, що акустична база ізольованих слів враховує кількість реалізацій кожної фонеми. Також наявність коротких синтагм сприяє покращенню результатів розпізнавання, оскільки саме короткі синтагми найбільш характерні для людського мовлення.

Таблиця 1 – Показники фонемної помилки розпізнавання *PER* (%) для КВ злитого мовлення на основі різних мовленнєвих образів, використовуючи різні акустичні моделі (злите мовлення – ЗМ та ізольовані слова – ІС) інструментарієм *HTK*

Назва КВ	Фонема		Відкритий склад		Склад за правилами складоподілу	
	Акустична модель, що використовується					
	ЗМ	ЗМ + ІС	ЗМ	ЗМ + ІС	ЗМ	ЗМ + ІС
«Випадкова» КВ	28,86	25,6	24,92	23,06	24,54	21,34
«Випадкова» КВ (без наголосу)	21,39	20,96	17,68	17,00	17,29	18,36
«Частотна» КВ	36,6	26,7	37,75	23,22	-	22,33
«Частотна» КВ (без наголосу)	26,1	21,56	27,95	17,49	-	18,11
КВ «Вікіпедія»	31,93	30,76	28,01	28,53	28,18	29,35
КВ «Вікіпедія» (без наголосу)	24,72	24,52	28,81	22,02	21,00	22,88

Акустичні моделі розпізнавання будувалися, враховуючи наголошеність голосних. На письмі наголос зазвичай не указується. У табл. 1 після розпізнавання було видалено інформацію про наголоси. Це штучним чином збільшило надійність розпізнавання на декілька відсотків.

А чи вплине на надійність розпізнавання, якщо в акустичній моделі не враховувати наголос голосних? Для дослідження цього була створена акустична модель, яка ігнорує ознаку наголошеності в алфавіті фонем.

Також, щоб врахувати специфіку українського вимовляння, а саме редукцію ненаголошених *е, и* до $е^{и}$, $и^{е}$ відповідно, була створена акустична модель, в якій були залишені тільки дві голосні *е́* та *и́*.

В табл. 2 наведені показники помилки розпізнавання *PER* (%) фонемного розпізнавання при використанні вищезгаданих акустичних моделей фонем.

Таблиця 2 – Показники фонемної помилки розпізнавання *PER* (%) без урахування наголошеності на різних акустичних моделях

Назва КВ	Акустична модель, яка використовувалася					
	ЗМ	ЗМ (наголос видалявся після розпізнавання)	ЗМ + ІС	ЗМ + ІС (наголос видалявся після розпізнавання)	ЗМ + ІС (без наголосу під час навчання)	ЗМ + ІС (наголошені тільки *е* та *и*)
«Випадкова» КВ	28,86	21,39	25,6	21,56	21,35	21,27
«Частотна» КВ	36,6	26,1	26,7	20,96	26,47	26,19
КВ «Вікіпедія»	31,93	24,72	30,76	24,52	21,43	21,75

Із табл. 2 випливає, що результати залежать як від способу формування акустичної моделі розпізнавання, так і від способу формування КВ. Результати, отримані на моделях тільки злитого мовлення, значно покращуються при об’єднані з моделями, побудованими на ізольованих словах. Слід зазначити, що результати досліджень, наведених у табл. 2, проводилися тільки для фонемного розпізнавання, – інші мовленнєві образи не використовувалися.

Ряд експериментів було проведено накладанням обмежень на послідовності елементів, застосовуючи лінгвістичну модель на фонемно-морфемному рівні. У табл. 3 наведені

результати пофонемного та поскладового розпізнавання різних КВ, застосовуючи біграмні лінгвістичні моделі для фонем та складів. Ці лінгвістичні моделі будувалися на початковому текстовому корпусі, при цьому розрізнялися наголошені та ненаголошені фонеми.

Таблиця 3 – Показники фонемної помилки розпізнавання *PER* (%) для КВ злитого мовлення на основі різних мовленнєвих образів, використовуючи біграмні лінгвістичні моделі інструментарієм *HTK*

Назва КВ	Фонема	Відкритий склад	Склад за правилами складоподілу
«Випадкова» КВ	24,80	27,52	*24,76*
«Випадкова» КВ (без наголосу)	18,22	21,26	*17,03*
«Частотна» КВ	27,68	24,16	*22,00*
«Частотна» КВ (без наголосу)	20,05	17,40	*15,34*
КВ «Вікіпедія»	28,23	31,85	*28,16*
КВ «Вікіпедія» (без наголосу)	*21,28*	25,06	21,34

Порівнюючи результати розпізнавання, наведені у табл. 1 та у табл. 3, можемо зробити висновок, що накладання обмежень через лінгвістичну модель, навіть таку, що побудована на відносно невеликій кількості текстів, дає покращення надійності розпізнавання (напівжирний шрифт – покращення відносно показників табл. 1, похилений шрифт – найкращий результат для даної КВ). Застосування фонемно-морфемної ЛМ для «Частотної» КВ дали найкращі показники *PER* (%). Для КВ «Вікіпедія» покращення не суттєві, оскільки, як ми припускаємо, при формуванні цієї контрольної вибірки не застосовувався початковий текстовий корпус, тому сформовані статистичні поскладові моделі мають незначний вплив на розпізнавання.

Адаптація системи розпізнавання мовлення до різних платформ

В рамках Державної науково-дослідницької програми «Образний комп'ютер» була розроблена низка прототипів мобільних пристроїв, на яких реалізовані технології і алгоритми розпізнавання та синтезу мовленнєвих сигналів. Вся лінійка мобільних пристроїв (цифровий диктофон, голосовий секретар та мобільний телефон) розроблялися на основі сигнальних процесорів *Analog Devices* сімейства *BlackFin*. Для цих процесорів існує можливість запуску на них операційного середовища *uCLinux*, яке належить до сімейства *UNIX*-подібних операційних систем, та базуються на вихідних кодах ядра ОС *Linux*. Використовуються три основних модуля – *GNU Toolchain* (крос-компілятор), *Das U-boot* (вихідні файли завантажувача), *Linux Kernel* (вихідні файли ядра ОС *uCLinux*).

GNU Toolchain (крос-компілятор) – спеціальний компілятор, який працює в операційному середовищі *Linux* на персональному комп'ютері і формує виконуваний код для операційного середовища *uCLinux* на основі сигнального процесора *AD BlackFin*. Цей компілятор використовується як для крос-компіляції вихідних кодів ядра ОС *uCLinux* і завантажувача *Das U-boot*, так і для крос-компіляції модулів, написаних мовою програмування *C\C++*, для можливості їх виконання в середовищі *uCLinux*. До складу компілятора входять такі основні модулі: компілятор *gcc* і *gcc-elf* (версії 3.4 і 4.1, що дає широкі можливості сумісності програм) і спеціалізована бібліотека для вбудовуваних

систем *uclibc*. Для зручності крос-компілятор *GNU Toolchain* надається у двох видах – у вигляді пакету *rpm* та у вигляді архівів *tar.gz*. Також крос-компілятор існує в 2-х версіях – для 32-бітних та для 64-бітних систем відповідно. Після встановлення пакетів весь функціонал буде доступний для використання стандартних методів процедури *make.*

Das U-boot (завантажувач) – комп’ютерний завантажувач операційних систем, орієнтований на вбудовані пристрої архітектур *MIPS*, *ARM* та інших. Після крос-компіляції може бути записаний у *Flash-ROM* платформи. Після чого код завантажувача виконується при запуску системи, що дає змогу завантажити в пам’ять та запустити ядро ОС *uCLinux*.

Linux Kernel (ядро ОС *uCLinux*) – центральна частина операційного середовища *uCLinux*, забезпечує різним процесам координований доступ до ресурсів комп’ютера, таким як процесорний час, оперативна пам’ять та зовнішнє апаратне забезпечення, та реалізовує функції файлової системи.

При розпізнаванні окремих слів система розпізнавання оперує тільки словником та акустичними моделями із бази даних та знань. Розпізнавання злитого мовлення потребує підключення породжувальних граматик у формі Бекуса-Наура або статистичної лінгвістичної моделі [2]. В останньому випадку декодер на початку використовує біграми, далі у вузлах сформованого графа динамічного програмування уточнюються значення часткової міри схожості із залученням N-грам, $N > 2$.

Декодер реалізований мовою програмування *C* на основі [10] для персонального комп’ютера та адаптований для можливості крос-компіляції до мікропрограмного коду операційного середовища *uCLinux* сигнального процесора *BF*-561.

Результати розпізнавання ідентичних фрагментів мовлення на ПК та на портативних пристроях збігаються з точністю до 6-го знаку після коми. Уніфікація програмного коду дає змогу всі дослідження проводити на персональному комп’ютері.

Також була розроблена система, яка реалізує клієнт-серверну ідеологію. При цьому клієнтські програми розроблялися мовою *Java* для найбільш розповсюдженої мобільної платформи *Android*. У клієнтському ПО реалізована можливість запису мовленнєвого сигналу, що розпізнається, набору тексту для подальшого озвучення та обміну інформацією зі своїми серверами. Серверне ПО розроблялося мовою *PHP* (отримання даних від клієнтів) та *C*++ (розпізнавання та синтез мовленнєвих сигналів). Обмін даними між клієнтом та сервером відбувається за протоколом *http* за допомогою стандартних процедур *POST* та *GET*. Обсяг словника обмежується обчислювальними можливостями сервера.

Обидва описаних підходи реалізації розпізнавання мовленнєвих сигналів на мобільних пристроях – ізольований та клієнт-серверний – закладають передумови до введення гібридного підходу, в якому передбачається спроба розпізнати мовлення безпосередньо на мобільному пристрої, а у випадку відмови розпізнавання – скористатися зв’язком із сервером.

Висновки

Проведений аналіз показав, що при адаптації модуля декодера до архітектури портативних пристроїв найбільш гнучкою є гібридна архітектура, яка дає змогу використати одночасно переваги ізольованого та клієнт-серверного підходів.

Підтверджено, що використання інформації про наголос є доцільним при формуванні акустичної моделі. Введення статистичних обмежень на порядок слідування елементів на фонемно-морфемному рівні дає змогу в цілому підвищити надійність

пофонемного розпізнавання. Це наближає перспективу реалізації багаторівневої моделі перетворення мовлення на текст.

Опрацьовано середовище для розроблення систем розпізнавання та синтезу мовлення на різних платформах. Це дасть змогу ефективно розробляти та випробовувати відповідне програмне забезпечення.

Подальші дослідження планується присвятити моделюванню взаємовпливу звуків у потоці мовлення, індивідуалізації параметрів моделі розпізнавання та інтеграції з технологією озвучення текстів. Вплив ряду параметрів декодеру на надійність і швидкість також є предметом майбутніх досліджень.

Литература

1. Винцюк Т.К. Анализ, распознавание и смысловая интерпретация речевых сигналов / Винцюк Т.К. – Киев : Наукова думка, 1987. – 263 с.
2. Gales M. The Application of Hidden Markov Models in Speech Recognition / M. Gales, S. Young // Foundations and Trends in Signal Processing. – 2007. – № 1(3). – P. 195-304.
3. Vintsiuk T. Multi-Level Multi-Decision Models in ASR / T. Vintsiuk, M. Sazhok // Proc. of the 10th International Workshop «Speech and Computer» (SPECOM'2005). – Patras, 2005. – P. 69-76.
4. [Електронний ресурс]. – Режим доступу : http://www.creaceed.com/ceedvocal/about
5. [Електронний ресурс]. – Режим доступу : http://www.apple.com/ios/siri/
6. [Електронний ресурс]. – Режим доступу : http://www.nuance.com/for-business/by-product/automotive-products-services/vocon-hybrid/
7. Робейко В.В. Багатозначна багаторівнева модель перетворення орфографічного тексту на фонемний / В.В. Робейко, М.М. Сажок // Штучний інтелект. – Донецьк, 2011. – № 4. – С. 117-125.
8. Сажок Н.Н. Кластеризация слов при построении лингвистической модели для автоматического распознавания речевого сигнала / Н.Н. Сажок // Кибернетика и вычислительная техника : Межведомственный сборник научных трудов. – Вып. 170. – Київ, 2012. – С. 59-66.
9. Васильєва Н.Б. Використання граматик вільного порядку слідування фонем і складів для пофонемного розпізнавання / Н.Б. Васильєва // Штучний інтелект. – Донецьк, 2011. – № 4. – С. 80-86.
10. Lee A. Julius – an open source real-time large vocabulary recognition engine / A. Lee, T. Kawahara // Proc. European Conference on Speech Communication and Technology (EUROSPEECH). – 2001.– P. 1691-1694.
11. Розроблення програмно-апаратних засобів базового модуля усномовної комп'ютерної технології, що вбудовується в сучасні комп'ютерні системи, створення на їх основі високотехнологічних електронних виробів широкого застосування та здійснення заходів для їх впровадження в виробництво (ОК_2009_2) : звіт про НДР (заключний). МННЦІТіС НАН та МОН України / [Кер. Т.К. Вінцюк]. – Київ, 2010. – 149 с. – № ДР 0109U004244.

Literatura

1. Vintsiuk T.K. The analysis, pattern recognition and semantic interpretation of the speech signals. – Kiev : Naukova Dumka, 1987. – 263 p.
2. M. Gales, S. Young. The Application of Hidden Markov Models in Speech Recognition. // Foundations and Trends in Signal Processing. – 2007. – № 1(3). – P. 195-304.
3. T. Vintsiuk, M. Sazhok. Multi-Level Multi-Decision Models in ASR. // Proc. of the 10th International Workshop "Speech and Computer" (SPECOM'2005). – Patras, 2005. – P. 69-76.
4. http://www.creaceed.com/ceedvocal/about
5. http://www.apple.com/ios/siri/
6. http://www.nuance.com/for-business/by-product/automotive-products-services/vocon-hybrid/
7. Robeiko V.V., Sazhok M.M. Multi-level multi-decision model transformation orthographic text to phonemic// Artificial Intelligence. – № 4'2011. – Donetsk, 2011. – P. 117-125.
8. Sazhok N.N. Clustering words for construction of a linguistic model for the automatic recognition of the speech signal // Cybernetics and Computer Science: Interagency collection of scientific papers. – Issue. 170. – Kyiv, 2012. – P. 59-66.

9. Vasylieva N.B. Free phoneme and syllable order grammar application for continuous speech phoneme-by-phoneme recognition. Artificial Intelligence. – № 4'2011. – Donetsk, 2011. – P. 80-86.
10. A. Lee, T. Kawahara. Julius – an open source real-time large vocabulary recognition engine. // Proc. European Conference on Speech Communication and Technology (EUROSPEECH) – 2001.
11. The development of software and hardware base module of computer speech technology embedded in modern computer systems, the creation on their basis of high-tech electronic products widespread use and implementation of measures for their implementation in production (OK_2009_2): report of research work (final). IRTC IT&S. Supervisor Vintsiuk T.K. – Kyiv, 2010. № DR 0109U004244.

RESUME

N.B. Vasylieva, D.Ja. Fedoryn

Problems of Cross-platform Speech Recognition System Creation

This paper considers the problems associated with development of a speech recognition system on different computing platforms. Remote, on-board and hybrid system architectures are analyzed. The ways to adapt the speech recognition components for cross-platform realization are proposed.

Particular attention is given to the data and knowledge base forming for acoustic, phonetic and lexical levels. Speech recognition results are compared for a variety of acoustic models. Among them are: continuous speech based model with or without isolated word combining, with or without lexical stress consideration including specific features of certain Ukrainian vowels.

Relation between speech recognition acoustic and linguistic components is being modeled as well as spoken element selection has been investigated. Applied element order constraining statistical methods showed promising results for phonemes and two types of syllables even in condition of relatively small training data.

Aspects of decoder unification on the DSP microprocessor architecture including the possibility of speech signal remote processing are described.

Стаття надійшла до редакції 26.06.2013.